Kerk Krew

Statistics 101C Final Project

Mohini Vembu: mohiniv96@ucla.edu

Mrinalini Garg: garg.mrinalini@gmail.com

Kitu Komya: kitu.komya@gmail.com

Our Best Model

- **Algorithm**: xgboost whose minimum MSE was 1395373 on Kaggle
- **Variables**: 102 total in training set
  - All but 2 variables were recoded into dummy/binary variables
  - 85 variables based on original dataset
  - 17 variables based on outside data source (LAFD website)
- **Some parameters**:
  - eta: 0.4
  - gamma: 10
  - max_depth: 4
  - min_child_weight: 20
  - nrounds: 65

Dataset Transformation: Variables Used

- **Datasets**: training_dummy & testing_dummy
- **Variables**: training_dummy has 102 variables,
  from which 100 were recoded as dummy/binary variables

  - 4 variables for **year**
  - 1 (non-dummy) numeric variable for **dispatch sequence**
  - 11 variables for **dispatch status**
  - 40 variables for **unit type**
  - 3 variables for **PPE level** (EMS, non-EMS, NA)
  - 25 variables for **hour** (24 recreated from original time variable + NA)
  - 17 variables for **battalion** (outside data source from LAFD website)
  - 1 (non-dummy) numeric response variable for **elapsed time**

Dataset Transformation: Refining Dataframe

- We only used **complete cases** (no NAs in rows) of training_dummy which reduced our observations from 2774370 to 2315071
- Then from training_dummy, we further split it into 80% training and 20% testing data because of R's memory limitations
- Finally, we ran xgboost on training data and predicted on our testing data and chose model with lowest MSE

xgboost parameters

- list function
  - booster: **“gbtree”**
  - objective: **“reg:linear”**
  - eta: **0.4** (looked into range from 0.4 to 1)
  - gamma: **10** (looked into range from 0 to 50)
  - max_depth: **4** (looked into range from 4 to 40)
  - min_child_weight: **20** (looked into range from 5 to 30)
  - subsample: **0.5** (looked into range from 0.5 to 1)
  - colsample_bytree: **0.5** (looked into range from 0.5 to 1)
  - lambda: **1**
- xgb.train function
  - nrounds: **65**
  - print_every_n: **10**
  - early_stop_round: **5**
  - maximize: **F**

xgboost algorithm

- **Basic Idea**: tree-based model and a variant on gradient boosting machine
- **Advantages**:
  - Accurate/good results on most datasets
  - Tunable parameters
  - Regularization allows it to avoid overfitting
  - Enabled, internal cross validation
  - Efficient tree pruning
  - Reduces misclassification error from boosting method (builds upon boosting algorithm)
  - Almost 10 times faster than random forest (RF algorithms took us 6-12 hours to compute)

Other variables/algorithms we considered

- Variables
  - Grouping Dispatch Sequence into smaller factors by frequency
  - Grouping Unit Type by similar boxplot distributions
  - Grouping Unit Type by frequency
  - Creating new variable: # of distinct incident.ID since multiple same incident.IDs were present
  - Using Bureau instead of Battalion (but Bureau was too general)
  - Using area square mile, shape area, and shape length of each battalion (data from LAFD site)
- Algorithms
  - Boosting
  - Random Forests